Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Analysis of 329,942 SARS-CoV-2 records retrieved from GISAID database

Maria Zelenova [a,b,*,1], Anna Ivanova [a,1], Semyon Semyonov [a], Yuriy Gankin [a,**]

[a] Quantori, 625 Massachusetts Ave, Cambridge, MA, 02139, USA
[b] Mental Health Research Center, Kashirskoe Shosse 34, 115522, Moscow, Russia

ARTICLE INFO

Keywords:
SARS-CoV-2
Bioinformatics
GISAID
SNP
Pandemic
Clustering
Machine learning
Sequencing
Correlation coefficient matrix

ABSTRACT

Background: The SARS-CoV-2 virus caused a worldwide pandemic – although none of its predecessors from the coronavirus family ever achieved such a scale. The key to understanding the global success of SARS-CoV-2 is hidden in its genome.

Materials and methods: We retrieved data for 329,942 SARS-CoV-2 records uploaded to the GISAID database from the beginning of the pandemic until the January 8, 2021. A Python variant detection script was developed to process the data using *pairwise2* from the BioPython library. Sequence alignments were performed for every gene separately (except ORF1ab, which was not studied). Genomes less than 26,000 nucleotides long were excluded from the research. Clustering was performed using HDBScan.

Results: Here, we addressed the genetic variability of SARS-CoV-2 using 329,942 samples. The analysis yielded 155 SNPs and deletions in more than 0.3% of the sequences. Clustering results suggested that a proportion of people (2.46%) was infected with a distinct subtype of the B.1.1.7 variant, which contained four to six additional mutations (G28881A, G28882A, G28883C, A23403G, A28095T, G25437T). Two clusters were formed by mutations in the samples uploaded predominantly by Denmark and Australia (1.48% and 2.51%, respectively). A correlation coefficient matrix detected 160 pairs of mutations (correlation coefficient greater than 0.7). We also addressed the completeness of the GISAID database, patient gender, and age. Finally, we found ORF6 and E to be the most conserved genes (96.15% and 94.66% of the sequences totally match the reference, respectively). Our results indicate multiple areas for further research in both SARS-CoV-2 studies and health science.

1. Introduction

A virus that appeared in Wuhan in December 2019 was soon recognized as a coronavirus, a single-stranded positive-sense RNA virus belonging to a Coronaviridae family. First discovered in the 1960s, two Coronaviridae family members (CoV-229E and CoVOC43) did not present a global threat [1,2]. However, a Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV, 2002/2003) and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV, 2012) changed public opinion: SARS-CoV left ~8098 people infected and ~774 dead; MERS-CoV caused ~2494 infections, leading to ~858 deaths. The SARS-CoV-2 exceeded the predecessors, infecting more than 159,319, 384 people worldwide and causing more than 3,311,780 deaths by the 12th of May 2021 (reported by Ref. [3]. The World Health Organization declared a SARS-CoV-2 - related pandemic and public health emergency on the January 30, 2020 [4]., [5]). The worst outcomes of the COVID-19, a disease caused by SARS-CoV-2, are currently associated with old age (65 and older), male gender, smoking, and comorbidities such as diabetes, cardiovascular disorders, and hypertension [6]. At present, over a year and a half later, the reasons for SARS-CoV-2 high transmissibility are still elusive (Kaur et al., 2020) [7]. Studies of viral genome, its evolution, and its mutations are especially beneficial in understanding the viral changing pattern [2]. Since common knowledge of SARS-CoV-2 proteins' functioning, signaling pathways, protein-protein, and protein-host cell interactions keeps rapidly accumulating due to the novelty of the virus, there is an urgent need to explore the SARS-CoV-2 changes [8].

* Corresponding author. Quantori, 625 Massachusetts Ave, Cambridge, MA, 02139, USA.
** Corresponding author.
E-mail addresses: maria_zelenova@yahoo.com (M. Zelenova), anna.ivanova@quantori.com (A. Ivanova), semyon.semyonov@quantori.com (S. Semyonov), yuriy.gankin@quantori.com (Y. Gankin).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2021.104981
Received 19 August 2021; Received in revised form 17 October 2021; Accepted 22 October 2021
Available online 26 October 2021

1.1. Describing the viral sequence

SARS-CoV-2 genome was first sequenced in January 2020, a month after COVID-19 became a worldwide alert [53]., [9]. The genome consists of 29903 nucleotides (GenBank accession number MN908947). Its length and overall genetic contents carry little surprise since it has long been established that coronaviruses have ones of the largest genomes amid all RNA viruses (varying from ~26 to ~32 kb in length) (Kaur et al., 2020). Although many mutations have currently been found in the viral genome [8], only a small number of them are high-frequency: 119 SNPs exceed the 0.3% threshold, according to Ref. [11]. Based on the mutations, eight distinct viral clades had been reported by GISAID and twelve by Nextstrain by March 2021. Specific SARS-CoV-2 variants caused the most concern: a B.1.1.7 caused a travel ban in December 2020 because of its increased transmissibility [12]; a B.1.351 was thought to be more abundant in healthy young people and result in a more severe disease course in those cases [13]; a P.1 was presumed to be more infectious [14]. A recent B.1.617.2 (delta) variant struck India in March 2021 and quickly became the most reported variant [15]. Most frequently, mutations are found in SARS-CoV-2 sequences coding for spike (S) protein, RNA-dependent RNA polymerase (RdRp), and nucleoprotein (N). Despite a vast amount of knowledge accumulating daily, the exact consequences of most viral mutations are unknown [2]. Current updates on the positions and functions of viral regions are presented in Table 1. Although any results of genomic variation analysis obtained using a bioinformatic approach should be considered with caution until experimental confirmation [2,7], bioinformatics plays an important role in unraveling the viral mysteries. Overall, SARS-CoV-2 genome mutations are hypothesized to impact viral transmissivity, case fatality risk, and numerous other features. In this paper, we describe our research aimed at analyzing 329,942 viral FASTA sequences obtained from human hosts to observe mutational changes and explore the accompanying data. The present work analyzes concomitant mutations on a large scale for the first time, emphasizes the importance of GISAID database changes and provides thorough evaluation of the patient data suggesting multiple prospective grounds for both novel research and vaccine targets.

2. Materials and Methods

Data for 329,942 SARS-CoV-2 genomes isolated from human hosts were retrieved from the GISAID database, along with additional information (records from the December 24, 2019 until the January 8, 2021). Custom code for revealing insertions, deletions, and SNPs was used alongside the “pairwise2 local” tool (https://biopython.org/docs/1.78/api/Bio.pairwise2.html) from the BioPython library (Python version 3.7, BioPython version 1.78; https://biopython.org/). Alignments were done for every viral gene separately, except ORF1ab, which was not considered in the present research. Every gene was aligned to a reference sequence, and final positions were calculated on a reference genome (accession number MN908947.3) [19]. Genomic positions were retrieved from the UCSC genome browser (see Table 1). We used Pandas (version 1.2; https://pandas.pydata.org/), Matplotlib (version 3.3; https://matplotlib.org/), and Seaborn (version 0.11; https://seaborn.pydata.org/installing.html) to visualize the data. Cluster analysis was executed using HDBScan (version 0.8; https://hdbscan.readthedocs.io/en/latest/) and visualized with t-SNE (t-distributed stochastic neighbor embedding; sklearn version 0.23; https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html) (Fig. 1). Clustering was performed using data on SNPs and deletions whose frequency exceeded 0.3% in the present research. Based on that cut-off (0.3% or 989 records) clustering parameters search was performed. The clustering parameters that yielded a minimum number of clusters, subject to the condition of at least 989 records in one cluster, were determined as suitable for the research. Final clustering parameters were set as follows: “minimum cluster size” – “2000”, “minimum samples” – “5”, “cluster selection epsilon” – “0.5”, “cluster selection method” – “eom”, “metric” – “euclidean”. Only sequences more than 26,000 nucleotides long were included in the study since the smaller sequences did not allow us to correctly align all genes of interest. We performed data filtering using the following steps: 1) the genomes that were less than 93% similar to the reference sequence were excluded from further analysis (as they contained low-quality sequences) 2) if unidentified symbols were determined in the aligned gene, and their count was not equal to the count of SNPs, the sequence was included in further research 3) we determined the percentage of match between the reference sequence and the aligned gene 3) gene sequences were divided according to the % of the matched genomes: 100% match to a reference genome was required to consider a sequence highly conservative, more than 99% match - to consider it moderately conservative, alignments in a range from 99% to 93% match were marked as low conservative. As these cutoffs were determined experimentally and we considered all the viral genes separately, we were free from simply deleting all the records containing ambiguous/unidentified symbols (“N”, “Y” etc.). Instead, examining genes separately increased the number of sequences that could be used in the research. Statistical significance was measured using a *t*-test and Bonferroni correction (for two parameters – age and gender). The correlation was measured using the Pearson correlation coefficient.

Table 1
SARS-CoV-2 genes, their genomic positions, length, and function as assumed to date (functions according to NCBI Gene, [16,17,18]]].

Viral gene	Genomic position (According to UCSC Genome Browser)	Gene length	Presumable main function
ORF1ab	266–21555	21290	Codes for polyproteins PP1ab and PP1a which allow for viral replication, transcription, and other functions
S	21563–25384	3822	Provides cell entry
ORF3a	25393–26220	828	Activates the NLRP3 inflammasome; may contribute to virus replication and pathogenesis
E	26245–26472	228	Facilitate virion assembly within cells
M	26523–27191	669	
ORF6	27202–27387	186	Likely promotes viral replication
ORF7a	27394–27759	366	Likely interacts with immune cells
ORF7b	27756–27887	132	The structural component of the SARS-CoV-2 virion
ORF8	27894–28259	366	Downregulates MHC-I
N	28274–29533	1260	Packages viral genome inside the capsid
ORF10	29558–29674	117	Not identified

3. Results

By the January 8, 2021, the GISAID database had SARS-CoV-2 records deposited by 142 countries. Even though more than 329,000 records had been uploaded up until then, these data had limited research potential due to several significant problems. First, some of the uploaded sequences were dramatically smaller than the reference sequence (e.g., <5000 nucleotides) or contained an enormous (more than 7% of each gene of interest) number of ambiguous letters (Fig. 2 represents the sequence size range obtained for the data used in current research; the smallest sequences were mostly obtained by Sanger sequencing). Another weakness was the lack of automation/control in terms of data entry to the system. That drawback led to numerous misspellings and data variants, along with missing information. Thus, the “collection date” field could include a year, a month, and a date, contain only the year, or, for some records, have a wrong year (e.g., 2002 instead of 2020). “Gender” and “Patient age” parameters were filled only for

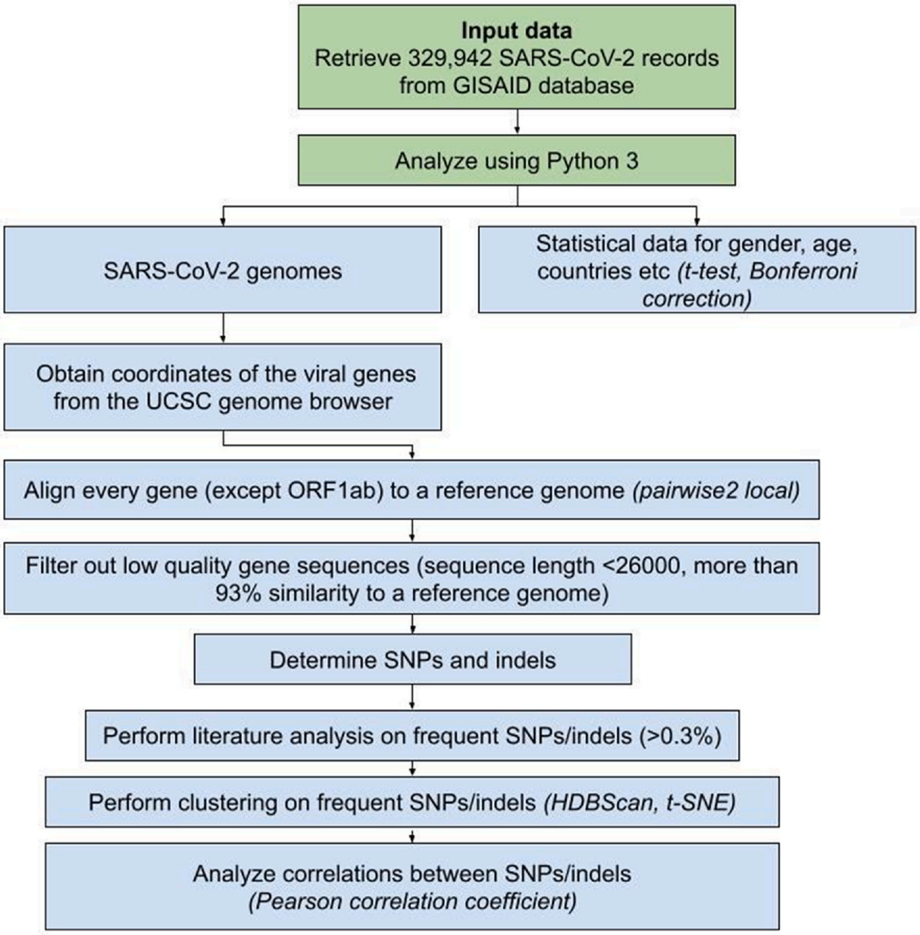
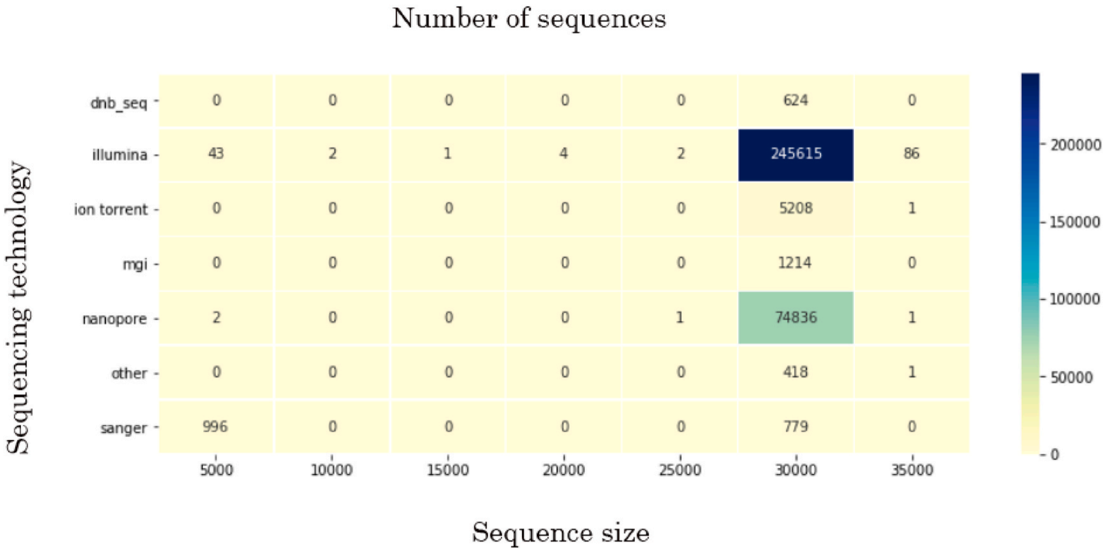

Input data
Retrieve 329,942 SARS-CoV-2 records from GISAID database
Analyze using Python 3
SARS-CoV-2 genomes
Statistical data for gender, age, countries etc (t-test, Bonferroni correction)
Obtain coordinates of the viral genes from the UCSC genome browser
Align every gene (except ORF1ab) to a reference genome (pairwise2 local)
Filter out low quality gene sequences (sequence length <26000, more than 93% similarity to a reference genome)
Determine SNPs and indels
Perform literature analysis on frequent SNPs/indels (>0.3%)
Perform clustering on frequent SNPs/indels (HDBScan, t-SNE)
Analyze correlations between SNPs/indels (Pearson correlation coefficient)

Fig. 1. Schematic representation of the methods used in the current work.



	5000	10000	15000	20000	25000	30000	35000
dnb_seq	0	0	0	0	0	624	0
illumina	43	2	1	4	2	245615	86
ion torrent	0	0	0	0	0	5208	1
mgi	0	0	0	0	0	1214	0
nanopore	2	0	0	0	1	74836	1
other	0	0	0	0	0	418	1
sanger	996	0	0	0	0	779	0

Fig. 2. The sequence size ranges obtained for the data used in current research.

23.3% and 23.1% of the records, respectively. The least informative for research was “Patient status,” which was not only filled for just 6.9% but also contained hardly interpretable data. Records’ bias was another problem. The prevalent number of genomes was uploaded by the United Kingdom (45.3%), USA (18.3%), Denmark (6.7%), and Australia (5.1%), with other countries’ input ranging from 3 to less than 1% of all records. Mean age was determined as 48 (confidence intervals (95% CI): 47.8, 48.1). Although gender values for a studied cohort equaled 52% of males and 48% of females (95% CI: 0.51, 0.52), mean gender values in some countries significantly declined from these numbers. Most gender inequality among records was noted in Saudi Arabia (80% males among 446 gender-filled records, p « 0.001), Singapore (75% among 1584 gender-filled records, p « 0.001), and Bangladesh (68% among 586 gender-filled records, p « 0.001) in terms of male prevalence, and South Africa (64% of females among 2591 gender-filled records, p « 0.001), Lithuania (61% among 193 gender-filled records, p « 0.001) and Russia (57% among 1545 gender-filled records, p « 0.001) in terms of female prevalence. The highest mean age was revealed in records submitted by the United Kingdom (59.6, p « 0.001) and France (59.5, p « 0.001), the lowest – by United Arab Emirates (35.6, p « 0.001), Gambia (37, p « 0.001), Oman (37.3, p « 0.012) and Bangladesh (38.5, p « 0.001). Only the countries which submitted more than 100 parameter-filled records were mentioned above (for full data, see Supplement 1). The records' bias also affected the patients' status. Some countries presumably uploaded the records with predominantly one or another status (e.g., out of all records uploaded by Brazil, 40% contained patient status "Dead").

3.1. Genomic data

The data were considered for every viral gene separately, except ORF1ab, which was not considered in the present research. While filtering the data to include only good-quality sequences (Table 2), we encountered an obscure phenomenon concerning an ORF7b gene. Nearly 11,290 (out of 329,942) FASTA records were featured by a similar pattern consisting of 52 "N"s (for most, genomic coordinates: 27757–27808). Sixty percent of that data was obtained using Nanopore sequencing (although 22.7% of all the data was acquired by that sequencing technology). Besides sequencing technology, the problem could derive from a particular assembly method, more precisely – from choosing a wrong method or unsuitable parameters, such as k-mer size. "Assembly method" data were present in 45.9% of all records, while "sequencing technology" – in 99.9%). For records where sequences contained stretches with 52 "N"s, the "assembly method" was filled for 23.5%. Since we could not estimate the assembly method and its parameters, we investigated the most prevalent methods among records containing stretches of 52 "N"s. The further research was limited due to multiple variations created by manual system entry.

3.2. Conservation

Analyzing the conservation of the genes allowed us to get some insights into their importance for the virus and potential treatment (Table 3).

3.3. Insertions and deletions

No insertions with a frequency greater than 0.3% were found. Two deletions were identified in the S gene: 21765-ATACATG > A with 4.67% frequency and 21991-TTTA > T with 2.94% frequency.

Table 2
Number of records included in the research after data filtering, except for ORF1ab, which was not considered in the present research.

Gene	Number of records included in the research after data filtering	%
ORF1ab	NA	NA
S	306,821	92.99%
ORF3a	313,597	95.05%
E	326,054	98.82%
M	322,967	97.89%
ORF6	327,034	99.12%
ORF7a	296,602	89.9%
ORF7b	299,007	90.62%
ORF8	320,383	97.1%
N	315,208	95.53%
ORF10	320,577	97.16%

Table 3
Conservation of viral genes.

Viral gene	Highly conservative, %	Moderately conservative, %	Low conservative, %
ORF1ab	NA	NA	NA
S	3.15	81.02	10.91
ORF3a	52.78	43.22	0.72
E	94.66	4.62	0.1
M	62.4	36.36	0.42
ORF6	96.15	3.13	0.25
ORF7a	83.43	7.12	0.62
ORF7b	85.48	5.47	0.4
ORF8	64.71	33.09	0.38
N	19.62	76.35	0.72
ORF10	73.87	23.62	0.16

3.4. SNPs

We analyzed genomic data with respect to the date of their upload, which allowed us not merely to determine the most frequent mutations but also to reveal and visualize their changes through the year (Supplement 2 contains data on SNPs occurring with more than 0.3% frequency among 329,942 viral genomes. Supplement 3 contains charts representing changes by month for each mutation).

3.5. Clustering

We applied HDBScan to the data on SNPs and deletions with a frequency greater than 0.3%, which resulted in 43 clusters (Fig. 3). Some data did not fit any cluster. A number of the forty-three clusters presented interesting data. Cluster #0 (size regarding all studied genomes - 1.77%) contained all mutations from a "British variant", except an SNP in the M gene (ORF1ab mutations were not considered due to the specificity of the research), in 100% records of the cluster. Four mutations were present in the cluster with 100% frequency - G28881A, G28882A, G28883C, and A23403G. Cluster #1 contained 0.69% of all records, had the mutations mentioned above (from the "British variant") and the following variants: A28095T (frequency in the cluster - 49.98%), G28881A, G28882A, G28883C, A23403G (100% each), and G25437T (31.58%). Cluster #20 showed significantly different parameters in terms of age and gender. The cluster included one mutation in ORF3a (G26144T) and was characterized by a mean age of 57 and a gender ratio of 50.46 males per 49.54 females. Cluster #25 was featured by the increased mean age (53) and could be described by 5 mutations occurring with different frequencies: A23403G (99%), G25563T (87%), C27964T (87%), C28977T (10%), and C23731T (2%). Cluster #34 demonstrated a decreased mean age of 43 and was represented by 9 mutations: C28869T (100%), C27964T (100%), A23403G (100%), G25563T (100%), G25907T (100%), C28472T (99%), G29402T (23%), A22255T (17%), G23593T (4%). Two clusters, #13 and #39, showed an altered male to female ratio. Cluster #13 was featured by 54.8% of males and 3 mutations: A23403G (100%), G25563T (100%), C26735T (5%); cluster #39 was characterized by 46.31% of males and 8 mutations: A23403G (99%), G22992A (99%), G23401A (99%), G28881A (99%), G28882A (99%), G28883C (99%), C27059T (7%), C22480T (6%). Mutations found in samples uploaded mainly by Denmark and Australia formed two clusters, each containing 8 mutations (sizes regarding all studied genomes - 1.48% and 2.51%, respectively): C26735T (100%), T26876C (100%), G25563T (100%), C25710T (100%), G29399A (100%), A23403G (99%), G22992A (99%), C27434T (13%) and A23403G (99%), G22992A (99%), G23401A (99%), G28881A (99%), G28882A (99%), G28883C (99%), C27059T (7%), C22480T (6%), respectively.

3.6. Concomitant mutations

According to a correlation coefficient matrix, 69 mutations had

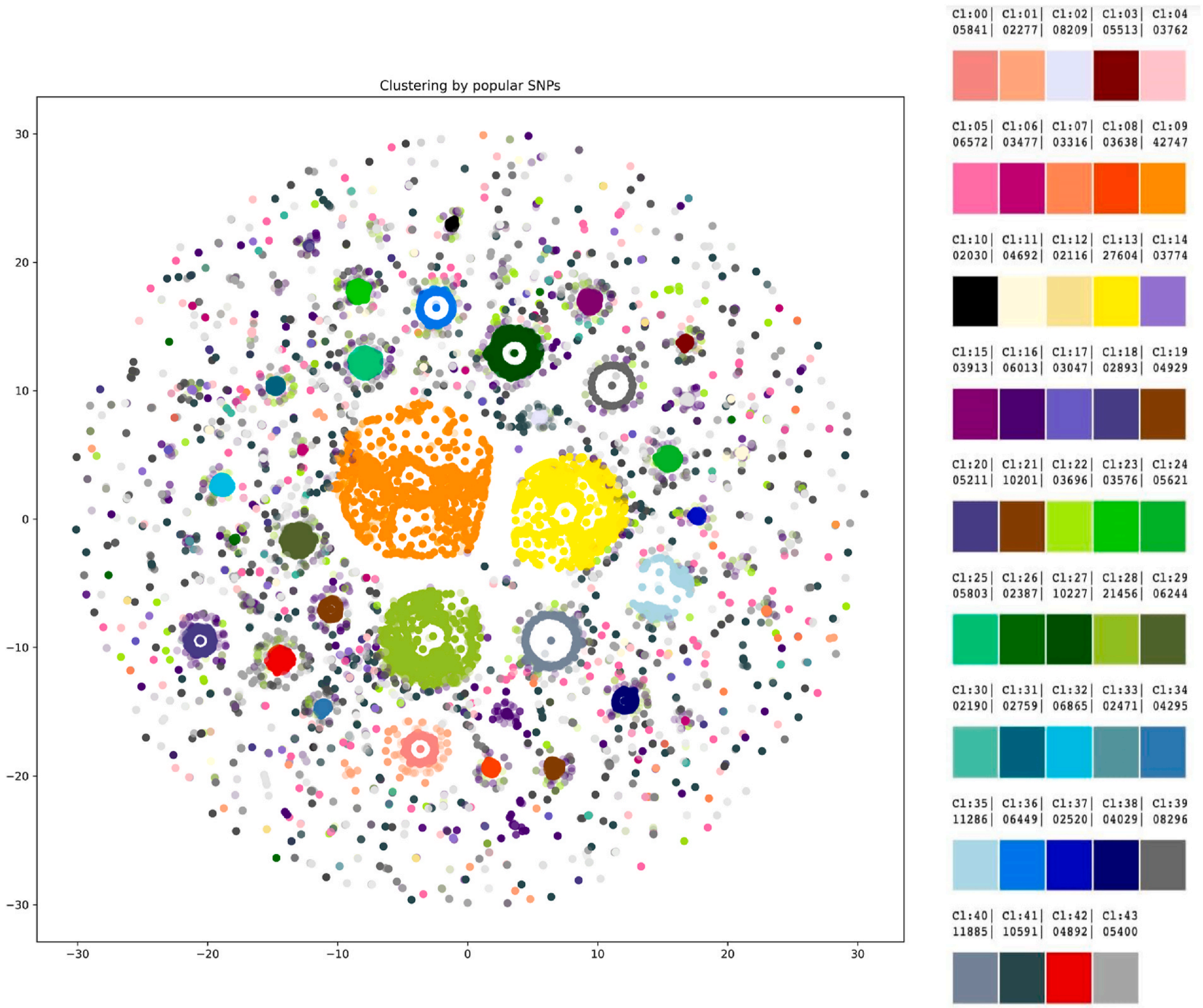


Fig. 3. Forty-three clusters were revealed by HDBScan. Legend on the right contains cluster numbers and color schemes.

correlations with at least one other mutation (Fig. 4; larger resolution and lower cutoff may be found in Supplement 4). In total, 160 pairs with a correlation coefficient greater than 0.7 were found (Supplement 5).

4. Discussion

The statistical and bioinformatic analysis of 329,942 records obtained from the GISAID database yielded data concerning many areas, from database design and medical care issues to genomic mutations and their probable effects. The abovementioned results are discussed below.

4.1. *Treatment targets: conservative sites*

At the moment, one of the most promising treatment and vaccine targets is the S protein, which enables the virus to enter human cells and is already targeted in such vaccines as Gam-COVID-Vac (Sputnik V), Oxford/AstraZeneca, Pfizer/BioNTech, and Moderna (Dai et al., 2020) [20]. However, the S gene has dramatically changed since the reference genome was first published – only 3.15% of the analyzed sequences totally match the reference sequence. Viral genes that changed least during the pandemic are ORF6 and E (96.15% and 94.66% of the sequences have 100% match the reference sequence, respectively). Although E protein acts together with an M protein in order to accomplish a virion assembly within the cells [21], the gene has changed dramatically less compared to M (62.4% of the sequences are highly conservative). According to these data, ORF6 and E are highly prospective targets for treatment/vaccine development. Currently, the E gene is only used as one of two qRT–PCR targets in SARS-CoV-2 detection assays by Roche (cobas® SARS-63 CoV-2 test). However, it is already known that the E protein of SARS-CoV-2 is highly immunogenic [22,23]. Researchers have attempted drug discovery concerning both E and ORF6. One group determined a drug-binding site of E's transmembrane domain using a solid-state nuclear magnetic resonance spectroscopy [24]. ORF6 can suppress both primary interferon production and interferon signaling. It is thought that SARS-CoV-2 with deleted ORF6 may be discussed in terms of intranasal live-but-attenuated vaccine invention [25]. Since ORF6 is one of three proteins causing the highest toxicity when overexpressed in human 293 T cells, and it also interacts with nucleopore proteins (RAE1, XPO1, RANBP2, and nucleoporins), treatment with an XPO1 inhibitor, Selinexor, was considered. Selinexor was found to reduce ORF6-induced toxicity in human 293 T cells [26]. Other groups found that Gliclazide and Memantine may inhibit E protein's channel activity, and Belachinal, Macaflavanone E, and Vibsanol B may inhibit the protein's function [27]; Gupta et al., 2020) [28].

4.2. *Clustering*

It may be proposed that, according to clustering results, although B.1.1.7 mutational contents may not be expanded due to the absence of the concomitant mutations in the general cohort, there is a proportion of people who got infected with its distinct subtype. The subtype may be

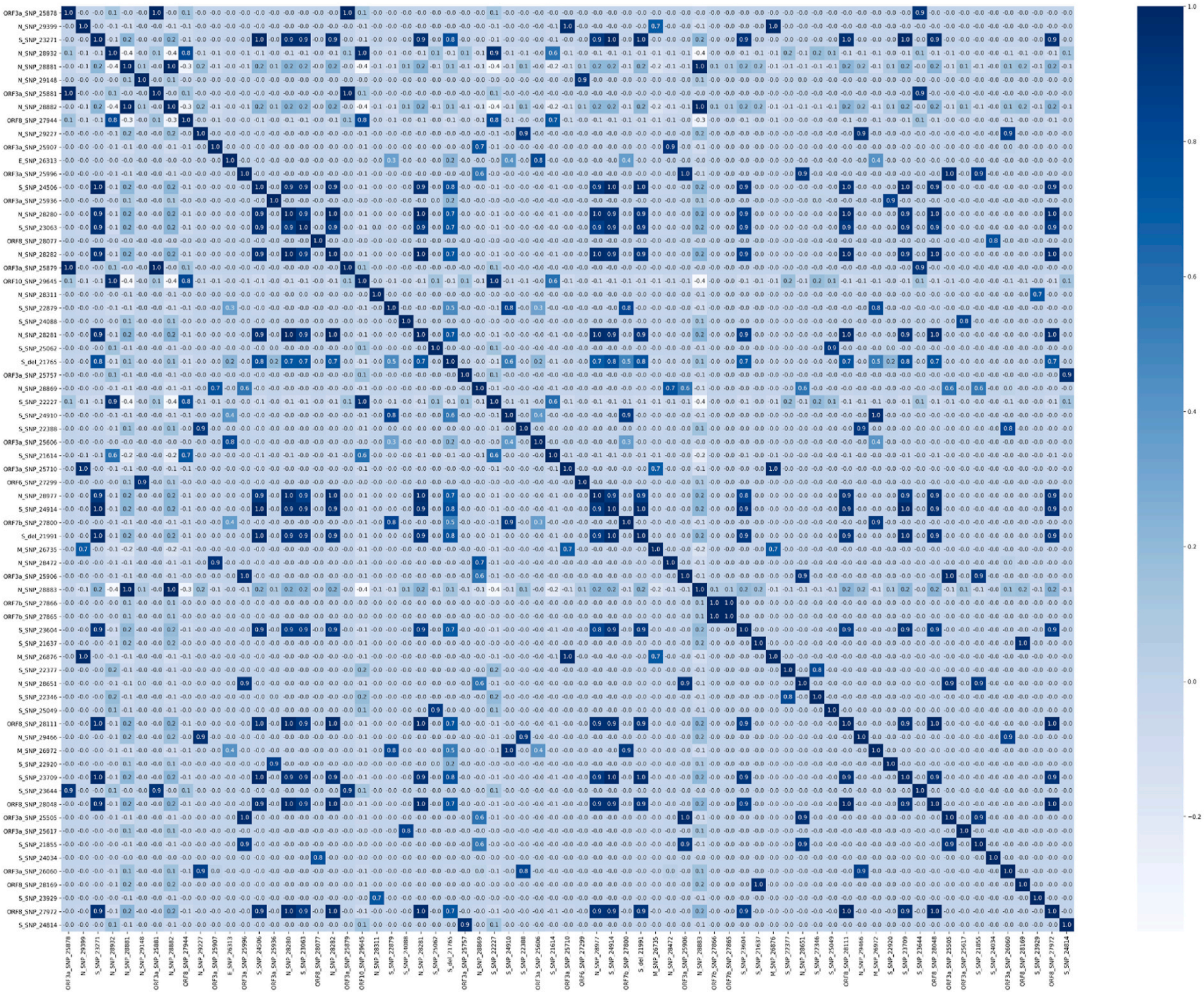


Fig. 4. Correlation coefficient matrix based on mutations with a frequency greater than 0.3%.

characterized by four to six additional mutations, with four being a more frequent option (G28881A, G28882A, G28883C, A23403G, A28095T, G25437T). Both clusters containing the "British variant" mutations were also the most recent, with a mean upload time of the middle of November 2020. A mutation in ORF3a (G26144T) that formed a cluster featured by increased age (57) and significantly different male to female ratio (50.46:49.54) has presumably disappeared from the population and was last noted in the uploads in September 2020. Due to increased age among patients carrying the virus with the mutation, it may be proposed to have increased virulence. Two clusters were associated with significantly different mean patient age (57 and 43), while two other clusters were featured by shifted male:female ratio: increased proportion of males in one (54.8%), and females – in the other (46.31%). Whether people of certain gender or age can be more prone to specific combinations of mutations is nevertheless unclear, and more research is needed in that direction. Mutations in samples uploaded predominantly by Denmark and Australia formed distinct clusters (8 mutations in each), which lets us speculate on the existence of so-called "Danish" and "Australian" variants.

4.3. Concomitant mutations

Current research shows that some mutations often present together with one or more others. In total, 160 pairs of mutations with a correlation coefficient greater than 0.7 were found. Most studies in this direction focus on certain concomitant mutations. For example, D614G is often considered together with P323L. Some researchers suggest the inability of D614G to cause viral success when presented alone [8,29]. T85I is noted to co-occur with Q57H, and P504L – with Y541C [8]. Also, R203K and G204R in the N gene were found to occur together with high frequency [30], which is confirmed in our research. G28881A is concomitant with G28882A and G28883C (r = 0.998). Variants of concern (e.g., B.1.1.7, B.1.351, P.1) also contain co-occurring mutations. However, to our knowledge, there are no publications analyzing concomitant mutations on a large scale. Therefore, our work shows this subject as a potentially fruitful ground for novel research.

4.4. The most frequent mutations

The most frequent mutation in the analyzed genes was a mutation in the S gene - A23403G (D614G), which was found in 94.15% of all studied genomes and in 99.9% of genomes uploaded in December 2020. D614G is considered to be more infectious than the ancestral form but not associated with increased disease severity [31]., [32,51]. Mutations with more than 20% frequency were found in different genes. In S, it was C22227T (A222V) with 22.25%. It was found in 53.8% of all uploaded sequences in November 2020 and assumed to influence viral transmissivity and antigenicity [33,34], as well as enhance the ability of the protein to interact with the environment [35]. A frequent mutation was also present in the M gene - C26801G (L93L) was observed in 21.82% (and 53.4%–43.2% of all uploads from November–December 2020). The assumed consequences of the mutation are yet to be described. The ORF3a gene had a G25563T (Q57H) mutation, found in 21.41% of the genomes. Four mutations with a frequency greater than 20% featured the N gene: G28881A (R203K), G28882A (R203R), G28883C (R203R), and C28932T (A220V). Interestingly, Q57H and R203K were found to cause substantial changes in protein structures (RMSD ≥5.0 Å). The mutations are also thought to affect the binding affinity of intraviral protein interactions [36]. Last, one most frequently occurring mutation found in ORF10, G29645T (V30L), was present in 22.03% of uploads in a general group and 44.6% of all uploads from December 2020. At the moment, it is proposed that ORF10 may not be a protein-coding gene, with its premature termination not affecting viral fitness or transmissivity [37].

4.5. Disappearing mutations potentially decrease viral fitness

Only three mutations have not been noted in the uploads for some time: G26144T (G251V) and G25979T (G196V) in ORF3a, which were last uploaded around September 2020 and early December 2020, respectively, and a C28836T (S188L) in the N gene, which was last seen around early to middle November. G251V results in the loss of a phosphatidylinositol-specific phospholipase X-box domain and a creation of a serine protease cleavage site [38]. Another work states that G251V and G196V might influence virulence, infectivity, ion channel activity, and viral release [39]. Might disappearing mutations impact viral fitness or human survival? The data is yet incomplete. However, in the present research G26144T (G251V) was found to create a cluster on its own; the mutation was featured by increased age (57) and an increased proportion of women compared to the general cohort.

4.6. Novel mutations

The most recent mutation in the current analysis is A28111G (Y73C) in ORF8, which appeared in the uploaded data about early September 2020. The mutation is included in a B.1.1.7 mutations' list. In total, B.1.1.7 is featured by 23 mutations [40] and is preliminarily reported as possibly associated with an increased risk of death [41]. We detected 13/14 mutations not located in the ORF1ab region and associated with the variant in the analyzed data. A T26801C mutation in the M gene was not found among mutations with a frequency greater than 0.3%, but our data yielded two mutations in the same position (freq >0.3%): C26801G and C26801T. The discrepancy could occur due to the differences in the reference sequences, which cannot be verified as Rambaut et al. did not specify the reference sequence number. We have also considered two other variants that have appeared lately - B.1.351 (a variant from South Africa) and P.1 (a variant from Brazil), but out of 8 and 14 non-ORF1ab mutations, respectively, only 2 and 3 were detected in our analysis among highly-present mutations. Consequently, it can be speculated that either a "British variant" has more transmissivity compared to the other two variants, or this result is due to a bias because of the number of the uploads.

4.7. GISAID database drawbacks lead to its severely limited research value

We have revealed that the major drawback of letting the users manually fill the fields of the records led to a loss of approximately 77%–93% of the data, depending on the parameter. The absence of quality control for genomic data yielded a presence of many sequences significantly shorter or longer than the reference genome (ranging from <5000 to 34000 nt). Many laboratories uploading the data did so significantly later than the sample collection date, some even a year later, which could distort the bioinformatic analysis. Certain laboratories indicated a month and a year, or only a year, of sample collection, omitting the day or day and month. An important analysis factor was that most data were uploaded by the United Kingdom, which created an overall data bias towards the UK statistics. As time is a crucial factor in a pandemic, a database update can be recommended in order to increase its value and quality.

4.8. Gender inequality in the uploaded data may reflect medical care availability issues

The cohort studied in the current research was represented by 52% of males and 48% of females (mean values; gender was not indicated for a subset of records). However, among records uploaded by Saudi Arabia, Singapore, and Bangladesh, men were present in 80%, 75%, and 68% of the records, respectively (while official statistics, male to female: Saudi Arabia - 58%:42%; Bangladesh 51%:49%, Singapore 52%:48%, by https://data.worldbank.org/). While Saudi Arabia is known for limiting access to medical care for women without a male guardian [42], Singapore, on the contrary, was ranked high (11th among 162 countries) for gender equality by the United Nations Development Programme last year [43]. The answer to this discrepancy most probably lies in the dormitories for migrants. In December of 2020, the Ministry of Health of Singapore declared that the majority of all COVID-19 cases occurred in migrant worker dormitories [44]. Although Bangladesh has shown significant improvement in moving towards gender equality (according to Ref. [45]), a medical access problem for rural areas persists. Estimating the rates of female inequality concerning medical care, a paper from the National Institute of Medical Health states that female patients were about half in number compared to male patients [46]. Our research also highlights possible issues in terms of health care for males: South Africa, Lithuania, and Russia uploaded 64%, 61%, and 57% of female records, respectively (the top three countries are considered for a shift in male to female ratio for both genders; while official statistics, male to female: South Africa 49%:51%, Lithuania 46%:54%, and Russia 46%:54%, by https://data.worldbank.org/). There are no data on limited medical care options for men in South Africa, Lithuania, or Russia. Thus, it can be speculated that the current lack of male patients may derive from a strong idea of masculinity (e.g., men must be strong and health complaints mean weakness) [47]. One more explanation is that more people working in the areas related to abundant social contact (e.g., medicine, education) in these countries are women. We suppose that this distribution may also be considered in terms of hospitalization criteria and sex differences between distinct age groups, and therefore leave this question to be still open for discussion.

4.9. Gender and age-related mutations

Although mean age across gender-filled records in our cohort was determined as 48 and mean gender as 52% of males and 48% of females, some mutations were characterized by increased or decreased age and shift in male to female ratio. A G23311C (E583D) was predominantly uploaded by the UK (97.1%), so it may be considered with respect to the other UK statistics. Among the records containing the SNP, the numbers (27% males and 73% of females) were obtained using 140 gender-filled records. In total, gender ratio among records uploaded by the UK (6275 records) was 50:50, however, for the current SNP, a solely UK number was 20:80, males to females (107 records). The patient age for the SNP was 61 (139 records), among only UK records – 68 (mean age in the UK was 59). We have not found literature data on the mutation with respect to age/gender. The only interesting message was an article stating that this mutation co-mutates with infectivity-enhancing S protein mutations, such as D614G, which cannot yet explain our finding [10].). Besides the aforementioned data, there were 12 mutations that were 10 points different in terms of gender and 2 – in terms of age. Due to the lack of data, only C23929T (Y789Y) and C28311T (P13L) could be considered further. P13L (mostly uploaded by Singapore in our research, 74% of males), is presumably associated with decreased deaths and significant changes altering the protein structure [19,48]. Age-related changes were noted for the mutations in the S (A22255T) and E (T26424C) genes, with characteristic ages of 38 and 62, respectively. For A22255T, 97.31% of the sequences were uploaded by the USA, and the total age-filled records' number for the SNP was 122, most uploaded by the USA. The mean patient age for the USA was 49. For a T26424C mutation, 97.96% of the sequences were uploaded by the UK, only 47 records were age-filled, most uploaded by the UK, where the mean patient age was 59. Increased age has been linked to the worst outcomes in those suffering from COVID-19. The mortality risk increases from 0 to 0.1% for children and adolescents under the age of 19 to 4.3–10.5% for the age group of 75–84 years. The most dramatic consequences are seen for individuals from 85 and older (up to 27.3% case fatality rate). Older patients get hospitalized more often (median age 74 vs median age of 43 for individuals in the outpatient care) and suffer from concomitant health issues (e.g., cardiovascular disorders, diabetes), which increase mortality rates by itself [32,49–51]. Interestingly, it has been repeatedly noted that men seem to suffer from COVID-19 more severely than women [52], with males proposedly being hospitalized more often than females (e.g. Refs. [32,51], report 67% of males versus 33% of females). Some mutations (for example, C27964T in ORF8) have been found to have gender dependence with a presumed ratio of 2:1 [8]. Although the reasons why males seem to be more severely affected are not yet clear, there are certain hypotheses on the topic. For instance, is it known that a primary way of SARS-CoV-2 entrance to the body is through its connection to angiotensin-converting enzyme 2 (ACE2), a part of the human renin-angiotensin-aldosterone system (RAAS) [53], and males show greater overall RAAS activity compared to females [54]. Also, as increased mortality risk is associated with cardiovascular diseases [55], the greater percentage of these disorders and thrombosis in men may contribute to fatality increase among males. A higher case fatality rate could also result from the fact that, in general, among intubated patients, men are more likely to acquire ventilator-associated pneumonia [56,57].

5. Conclusions

In this paper, we have analyzed 329,942 SARS-CoV-2 records obtained from the GISAID database. We addressed the quality of the uploaded records, gender distribution, gene conservation, SNPs, insertions and deletions, clusters, and a correlation coefficient matrix. Our research showed that mutations occurring with high frequency (>0.3%) were not abundant and constituted 155 changes concerning all genes (except ORF1ab, which was not considered in a current work). Many mutations presented with concomitant changes, which could alter their consequences for the virus or a human host. A large number of co-occurring mutations creates grounds for research on their meaning, as well as a probability of the occurrence in terms of novel mutations and concomitant variants. Conservation analysis suggested ORF6 and E genes as prospective treatment/vaccine targets due to their high conservation. Clustering allowed speculations on the existence of a subtype of a B.1.1.7 variant and the possible existence of variants specific to Denmark and Australia. Taken together, our results describe the genetic variability of SARS-CoV-2 and may be used for further research in different scientific areas.

Funding and grant disclosures

All authors are employed by the commercial company Quantori in Cambridge, Massachusetts, United States. Quantori provided support in the form of salaries for the employed authors.

Contributors

Maria Zelenova: contributed equally to this work with Anna Ivanova; performed experiment design, responsible for conceptualization, methodology, validation, writing – original draft, writing – review and editing.

Anna Ivanova: contributed equally to this work with Maria Zelenova; responsible for conceptualization, methodology, validation, visualization, writing – review.

Semyon Semyonov: responsible for methodology, software, data curation, project administration, resources, validation, writing – review.

Yuriy Gankin: an inspirer for the project due to COVID-19 situation; responsible for conceptualization, methodology, project administration, resources, supervision, writing – original draft, writing – review & editing.

All authors have approved the final article.

Summary

The present research paper analyses data for 329,942 SARS-CoV-2 records uploaded to the GISAID database from the beginning of the pandemic until the January 8, 2021. We addressed the quality of the uploaded records, gender distribution, gene conservation, SNPs, insertions and deletions, clusters, and concomitant mutations. To process the data, a Python variant detection script was developed, using *pairwise2* from the BioPython library. Current article shows that mutations occurring with high frequency (>0.3%) are not abundant and constitute 155 changes concerning all genes (except ORF1ab, which was not considered in a current work). Many mutations present with concomitant changes, which may alter their consequences for the virus or a human host. A large number of co-occurring mutations (160 pairs) creates grounds for research on their meaning, as well as a probability of the occurrence in terms of novel mutations and concomitant variants. Conservation analysis suggests ORF6 and E genes as prospective treatment/vaccine targets due to their high conservation (96.15% and 94.66% of the sequences totally match the reference, respectively). Clustering allows speculations on the existence of a subtype of a B.1.1.7 variant and a possible existence of variants specific to Denmark and Australia. The article also addresses the completeness of the GISAID database, patient gender and age differences. Taken together, our results describe the genetic variability of SARS-CoV-2 and may be used for further research in different scientific areas.

A conflict of interest statement

None Declared.

Declaration of competing interest

The authors declare there are no competing interests.

Acknowledgments

Authors thank Tatiana Tatarinova, Dallas Dorsey, and Isaiah Knox from the University of La Verne, Nika Tsutskiridze, Tsotne Khetsuriani, Revaz Mgeladze, Nona Kuloshvili, Tinatin Mekvabishvili from Quantori, Artem Artemov from the Medical University of Vienna, and Alexander Mikov from Amedart for insightful comments and discussions. We also thank Esther Alder from the Century College, Minnesota for text corrections.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104981.

References

[1] Furong Qi, Qian Shen, Shuye Zhang, Zheng Zhang, Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses, Biochem. Biophys. Res. Commun. 526 (1) (2020) 135–140.

[2] Changchuan Yin, Genotyping coronavirus SARS-CoV-2: methods and implications, Genomics 112 (5) (2020) 3588–3596.

[3] World Health Organization. https://covid19.who.int/. (Accessed 12 May 2021).

[4] Stefanie Weber, Christina Ramirez, Doerfler Walter, Signal hotspot mutations in SARS-CoV-2 genomes evolve as the virus spreads and actively replicates in different parts of the World, Virus Res. 289 (November) (2020), 198170.

[5] Alexandra C. Walls, Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, David Veesler, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, Cell 183 (6) (2020) 1735.

[6] Eric de Sousa, Dário Ligeiro, Joana R. Lérias, Chao Zhang, Chiara Agrati, Mohamed Osman, Sherif A. El-Kafrawy, et al., Mortality in COVID-19 disease patients: correlating the association of major histocompatibility complex (MHC) with severe Acute respiratory Syndrome 2 (SARS-CoV-2) variants, Int. J. Infect. Dis.: IJID: Off. Publ. Int. Soc. Infect. Dis. 98 (September) (2020) 454–459.

[7] Navpreet Kaur, Rimaljot Singh, Zahid Dar, Rakesh Kumar Bijarnia, Neelima Dhingra, Tanzeer Kaur, Genetic comparison among various coronavirus strains for the identification of potential vaccine targets of SARS-CoV2, Infect. Genet. Evol.: J. Mol. Epidemiol. Evolut. Genet. Infect. Dis. 89 (April) (2021), 104490.

[8] Rui Wang, Yuta Hozumi, Changchuan Yin, Guo-Wei Wei, Decoding SARS-CoV-2 transmission and evolution and ramifications for COVID-19 diagnosis, vaccine, and medicine, J. Chem. Inf. Model. 60 (12) (2020) 5853–5865.

[9] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, et al., A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (8) (2020) 727–733.

[10] Rui Wang, Jiahui Chen, Kaifu Gao, Yuta Hozumi, Changchuan Yin, Guowei Wei, Characterizing SARS-CoV-2 mutations in the United States, Res. Square (2020), https://doi.org/10.21203/rs.3.rs-49671/v1. August.

[11] Fangfeng Yuan, Liping Wang, Ying Fang, Leyi Wang, Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity, Transbound. Emerg. Dis. (2020), https://doi.org/10.1111/tbed.13931. November.

[12] Meera Chand, Susan Hopkins, Gavin Dabrera, Christina Achison, Wendy Barclay, Neil Ferguson, Erik Volz, Nick Loman, Andrew Rambaut, Jeff Barrett, Investigation of Novel SARS-COV-2 Variant: Variant of Concern 202012/01 (Report), Public Health England, 2020, p. 2.

[13] Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, et al., "Emergence and Rapid Spread of a New Severe Acute Respiratory Syndrome-Related Coronavirus 2 (SARS-CoV-2) Lineage with Multiple Spike Mutations in South Africa." *medRxiv*, 2020. https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1.full.

[14] Nuno R. Faria, Ingra Morales Claro, Darlan Candido, L.A. Moyses Franco, Pamela S. Andrade, Thais M. Coletti, Camila A.M. Silva, et al., Genomic characterisation of an emergent SARS-CoV-2 Lineage in Manaus: preliminary findings, Virological (2021). https://www.icpcovid.com/sites/default/files/2021-01/Ep%20102-1%20Genomic%20characterisation%20of%20an%20emergent%20SARS-CoV-2%20lineage%20in%20Manaus%20Genomic%20Epidemiology%20-%20Virological.pdf.

[15] Lopez Bernal, Jamie, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Thelwall Simon, Julia Stowe, et al., Effectiveness of covid-19 vaccines against the B.1.617.2 (delta) variant, N. Engl. J. Med. (2021), https://doi.org/10.1056/NEJMoa2108891. July.

[16] Lisa Miorin, Thomas Kehrer, Maria Teresa Sanchez-Aparicio, Ke Zhang, Phillip Cohen, Roosheel S. Patel, Anastasija Cupic, et al., SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling, Proc. Natl. Acad. Sci. U. S. A. 117 (45) (2020) 28344–28354.

[17] Ziliang Zhou, Chunliu Huang, Zhechong Zhou, Zhaoxia Huang, Lili Su, Sisi Kang, Xiaoxue Chen, et al., Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14+ monocytes, iScience 24 (3) (2021), 102187.

[18] Matthew D. Park, Immune evasion via SARS-CoV-2 ORF8 protein? Nat. Rev. Immunol. 20 (7) (2020) 408.

[19] Fan Wu, Zhao Su, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, et al., A new coronavirus associated with human respiratory disease in China, Nature 579 (7798) (2020) 265–269.

[20] Lianpan Dai, George F. Gao, Viral targets for vaccines against COVID-19, Nat. Rev. Immunol. 21 (2) (2021) 73–82.

[21] Philip V'kovski, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, Volker Thiel, Coronavirus biology and replication: implications for SARS-CoV-2, Nat. Rev. Microbiol. 19 (3) (2021) 155–170.

[22] Shreya Bhattacharya, Arundhati Banerjee, Sujay Ray, Development of new vaccine target against SARS-CoV2 using envelope (E) protein: an evolutionary, molecular modeling and docking based study, Int. J. Biol. Macromol. 172 (March) (2021) 74–81.

[23] Bruno Tilocca, Alessio Soggiu, Maurizio Sanguinetti, Gabriele Babini, Flavio De Maio, Domenico Britti, Alfonso Zecconi, Luigi Bonizzi, Andrea Urbani, Paola Roncada, Immunoinformatic analysis of the SARS-CoV-2 envelope protein as a strategy to assess cross-protection against COVID-19, Microb. Infect./Institut Pasteur 22 (4–5) (2020) 182–187.

[24] Venkata S. Mandala, Matthew J. McKay, Alexander A. Shcherbakov, Aurelio J. Dregni, Antonios Kolocouris, Mei Hong, Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers, Nat. Struct. Mol. Biol. 27 (12) (2020) 1202–1208.

[25] Chun-Kit Yuen, Joy-Yan Lam, Wan-Man Wong, Long-Fung Mak, Xiaohui Wang, Hin Chu, Jian-Piao Cai, et al., SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 Function as Potent Interferon Antagonists, Emerg. Microb. Infect. 9 (1) (2020) 1418–1428.

[26] Jin-Gu Lee, Weiliang Huang, Hangnoh Lee, Joyce van de Leemput, Maureen A. Kane, Zhe Han, Characterization of SARS-CoV-2 proteins reveals Orf6 pathogenicity, subcellular localization, host interactions and attenuation by selinexor, Cell Biosci. 11 (1) (2021) 58.

[27] Tomar Singh, Prabhat Pratap, Isaiah T. Arkin, SARS-CoV-2 E protein is a potential ion channel that can Be inhibited by Gliclazide and memantine, Biochem. Biophys. Res. Commun. 530 (1) (2020) 10–14.

[28] Manoj Kumar Gupta, Sarojamma Vemula, Ravindra Donde, Gayatri Gouda, Lambodar Behera, Ramakrishna Vadde, In-silico approaches to detect inhibitors of the human severe Acute respiratory Syndrome coronavirus envelope protein ion channel, J. Biomol. Struct. Dynam. 39 (7) (2021) 2617–2627.

[29] Sten Ilmjärv, Fabien Abdul, Silvia Acosta-Gutiérrez, Carolina Estarellas, Ioannis Galdadas, Marina Casimir, Marco Alessandrini, Francesco Luigi Gervasio, Karl-Heinz Krause, Epidemiologically Most Successful SARS-CoV-2 Variant: Concurrent Mutations in RNA-dependent RNA Polymerase and Spike Protein." *medRxiv*, 2020. https://www.medrxiv.org/content/10.1101/2020.08.23.20180281v1.abstract.

[30] M. Shaminur Rahman, M. Rafiul Islam, A.S.M. Rubayet Ul Alam, Israt Islam, M. Nazmul Hoque, Salma Akter, Md Mizanur Rahaman, Munawar Sultana, M. Anwar Hossain, Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences, J. Med. Virol. 93 (4) (2021) 2177–2195.

[31] Leonid Yurkovetskiy, Xue Wang, Kristen E. Pascal, Christopher Tomkins-Tinch, Thomas P. Nyalile, Yetao Wang, Alina Baum, et al., Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant, Cell 183 (3) (2020) 739–751, e8.

[32] Bette Korber, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al., Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus, Cell 182 (4) (2020) 812–827, e19.

[33] Emma B. Hodcroft, Moira Zuber, Sarah Nadeau, Katharine H.D. Crawford, Jesse D. Bloom, David Veesler, Timothy G. Vaughan, et al., Emergence and Spread of a SARS-CoV-2 Variant through Europe in the Summer of 2020." *medRxiv : the Preprint Server for Health Sciences*, 2020, https://doi.org/10.1101/2020.10.25.20219063. November.

[34] Barbara Bartolini, Martina Rueca, Cesare Ernesto Maria Gruber, Francesco Messina, Emanuela Giombini, Giuseppe Ippolito, Maria Rosaria Capobianchi, Antonino Di Caro, "The Newly Introduced SARS-CoV-2 Variant A222V Is Rapidly Spreading in Lazio Region, Italy." *medRxiv*, 2020. https://www.medrxiv.org/content/10.1101/2020.11.28.20237016v1.abstract.

[35] J.R. Lon, B. Xi, B. Zhong, Y. Zheng, P. Guo, Z. Chen, H. Du, Molecular Dynamics Simulation Study of Effects of Key Mutations in SARS-CoV-2 on Protein Structures." *bioRxiv*, 2021. https://www.biorxiv.org/content/10.1101/2021.02.03.429495v1.abstract.

[36] Siqi Wu, Tian Chang, Panpan Liu, Dongjie Guo, Wei Zheng, Xiaoqiang Huang, Yang Zhang, Lijun Liu, Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions, J. Med. Virol. 93 (4) (2021) 2132–2140.

[37] Katarzyna Pancer, Aleksandra Milewska, Katarzyna Owczarek, Agnieszka Dabrowska, Michał Kowalski, Paweł P. Łabaj, Wojciech Branicki, Marek Sanak, Krzysztof Pyrc, The SARS-CoV-2 ORF10 is not essential in vitro or in vivo in humans, PLoS Pathog. 16 (12) (2020), e1008959.

[38] Elio Issa, Georgi Merhi, Balig Panossian, Tamara Salloum, Sima Tokajian, SARS-CoV-2 and ORF3a: Non-synonymous Mutations and Polyproline Regions." bioRxiv, 2020, https://doi.org/10.1101/2020.03.27.012013.

[39] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, The species severe Acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nat. Microbiol. 5 (4) (2020) 536–544.

[40] Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L. Robertson, Erik Volz, On Behalf of COVID-19 Genomics Consortium UK (CoG-UK). Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations, 2020. https://virological.org. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.

[41] Dan Frampton, Tommy Rampling, Aidan Cross, Heather Bailey, Judith Heaney, Matthew Byott, Rebecca Scott, et al., Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in london, UK: a whole-genome sequencing and hospital-based cohort study, Lancet Infect. Dis. (2021), https://doi.org/10.1016/S1473-3099(21)00170-5. April.

[42]report World Report 2020. https://www.hrw.org/world-report/2020/country-chapters/saudi-arabia. (Accessed 20 March 2021).

[43]report Human Development Report - Singapore. http://hdr.undp.org/en/countries/profiles/SGPAccessed. (Accessed 20 March 2021).

[44]Measures to Contain the COVID-19 Outbreak in Migrant Worker D Measures to contain the COVID-19 outbreak in migrant worker dormitories. https://www.moh.gov.sg/news-highlights/details/measures-to-contain-the-covid-19-outbreak-in-migrant-worker-dormitories. (Accessed 12 May 2021).

[45]report Human Development Report 2020 – Bangladesh. http://hdr.undp.org/sites/all/themes/hdr_theme/country-notes/BGD.pdf. (Accessed 20 May 2021).

[46] Nazmun Nahar Nuri, Malabika Sarker, Helal Uddin Ahmed, Mohammad Didar Hossain, Fekri Dureab, Faith Agbozo, Albrecht Jahn, Overall care-seeking pattern and gender disparity at a specialized mental hospital in Bangladesh, Mater. Soc. Med. 31 (1) (2019) 35–39.

[47] Christopher J. Colvin, Gender, health and change in South Africa: three ways of working with men and boys for gender justice, Rech. Sociol. Anthropol. : RSC Adv. 48 (1) (2017) 109–124.

[48] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (10223) (2020) 497–506.

[49] Jennifer Beam Dowd, Liliana Andriano, David M. Brazel, Valentina Rotondi, Per Block, Xuejie Ding, Yan Liu, Melinda C. Mills, Demographic science aids in understanding the spread and fatality rates of COVID-19, Proc. Natl. Acad. Sci. U. S. A. 117 (18) (2020) 9696–9698.

[50] Daniel E.L. Promislow, A geroscience perspective on COVID-19 mortality, J. Gerontol. Ser. A, Biol. Sci. Med. Sci. 75 (9) (2020) e30–33.

[51] B. Korber, W. Fischer, S.G. Gnanakaran, H. Yoon, "Spike Mutation Pipeline Reveals the Emergence of a More Transmissible Form of SARS-CoV-2." *BioRxiv*, 2020. https://www.biorxiv.org/content/10.1101/2020.04.29.069054v2.abstract.

[52] P. Conti, A. Younes, Coronavirus COV-19/SARS-CoV-2 affects women less than men: clinical response to viral infection, J. Biol. Regul. Homeost. Agents 34 (2) (2020) 339–343.

[53] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Si Hao-Rui, et al., Addendum: a pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 588 (7836) (2020) E6.

[54] A.A. Zalucky, D.D.M. Nicholl, M.C. Mann, B.R. Hemmelgarn, T.C. Turin, J. M. Macrae, D.Y. Sola, S.B. Ahmed, Sex influences the effect of body mass index on the vascular response to angiotensin II in humans, Obesity 22 (3) (2014) 739–746.

[55] Jing Yang, Ya Zheng, Xi Gou, Ke Pu, Zhaofeng Chen, Qinghong Guo, Rui Ji, Haojia Wang, Yuping Wang, Yongning Zhou, Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis, Int. J. Infect. Dis.: IJID: Off. Publ. Int. Soc. Infect. Dis. 94 (May) (2020) 91–95.

[56] D.J. Cook, M.H. Kollef, Risk factors for ICU-acquired pneumonia, J. Am. Med. Assoc.: JAMA, J. Am. Med. Assoc. 279 (20) (1998) 1605–1606.

[57] Sofia B. Ahmed, Sandra M. Dumanski, Sex, gender and COVID-19: a call to action, Can. J. Publ. Health. Rev. Can. Sante Publ. 111 (6) (2020) 980–983.